Development and Evaluation of Audio-Visual ASR: A Study on Connected Digit Recognition

Michael T Chan
Rockwell Scientific Company
1049 Camino Dos Rios
Thousand Oaks, CA 91360
E-mail: mtchan@rwsc.com

Abstract

We present our findings from audio-visual speech recognition experiments for connected digit recognition in noisy environments. We derive hybrid (geometric- and appearance-based) visual lip features using a real-time lip tracking algorithm that we proposed previously. Using a small single-speaker corpus modeled after the TIDIGITS database, we build whole-word HMMs using both single-stream and 2-stream modeling strategies. For the 2-stream HMM method, we use stream-dependent weights to adjust the relative contributions of the two feature streams based on the acoustic SNR level. The 2-stream HMM consistently gave the lowest WER, with an error reduction of 83% at –3dB SNR level compared to the acoustic-only baseline. Visual-only ASR WER at 6.85% was also achieved. A real-time system prototype was developed for concept demonstration.

1. Introduction.

By combining acoustic and visual lip features for speech recognition, the resulting bimodal speech recognizer is markedly more robust in the presence of a variety of acoustic noise, when compared to the acoustic-only counterpart. The idea was pursued in a number of past studies [2][5][6][7][8][12][13][14][15][16][17][21]. Two key elements of an audio-visual speech recognition system are: (1) a front end for visual feature extraction, and (2) an information fusion architecture for integrating features from the two modalities. In recent years, considerable progress has been made in the first area [4][13][15][16], as well as in the second area [6][8][14][15][17].

There are primarily two categories of visual feature representation in the context of speech recognition. The first is model-based or geometric-based. Examples of such features are the width and height of the mouth (and their temporal derivatives) that can be estimated from the images using a tracking procedure. The second category is pixel-based or appearance-based; that is, the features are directly derived from the raw pixel values. The first category is more intuitive, but there is typically a substantial loss of information because of the data reduction involved. There is little loss of information in the second representation, but the high dimensionality of the image space is a computational disadvantage, and pixel-based features do not directly relate to observable articulator motion. Furthermore, normalization needed to account for lighting changes, translation and other effects is more difficult compared to the geometric-based counterpart.

We had experimented with a visual feature representation that combined the two types of features in our previous work and demonstrated its effectiveness in simple isolated digit recognition experiments [4]. The technique is adopted in the work reported in this paper. Here we develop new experiments to evaluate our system using stream-weighted 2-stream Hidden Markov Models (HMMs) as well as the traditional single stream HMMs in the context of connected digit recognition.

The rest of the paper is organized as follows. We first briefly describe our lip localization and tracking algorithms that allow geometric-based features to be extracted automatically, and pixel-based features to be subsequently normalized. We then focus on the proposed hybrid feature and its efficacy in the context of visual-only speech recognition. Finally, we describe the recognition experiments we performed, and report our findings from these experiments involving audio-visual speech recognition of connected digits in the presence of aircraft cockpit noise of varying SNR levels.

2. Visual Tracking and Localization.

To automate machine lipreading, we need to locate and track movements and appearance changes of the lips. Several model-based approaches for tracking lip movements that have been proposed include snake models [10], deformable templates [20], active shape models [12], and active contours [11]. We have developed an integrated approach addressing both lip localization and lip tracking [2][3]. The first part is based on Gaussian mixture model-based clustering using hue in the HSV color space. The largest elliptical connected region detected with the expected range of hue values is identified as the lips. It is usually quite effective and can be used to initialize the lip tracking part. Tracking is based on a user-specific 2D B-spline model that can be constructed offline, or estimated from sample images [3]. To optimize tracking stability, the model deforms only in an affine subspace, which is adequate for capturing most lip movements that occur in normal speech utterances. The model is driven (or fitted) based on locations of steepest gradient in the image, in a linearly transformed color space given by

$$s = \alpha \cdot r + \beta \cdot g + \gamma \cdot b,$$

where $\{\alpha, \beta, \gamma\}$ are speaker-dependent and are estimated based on linear discriminant analysis on the RGB content [3]. This overcomes problems associated with often fuzzy definition of lip boundary in the luminance channel, and the algorithm is consequently markedly more robust compared to most snake-based algorithms and other approaches based on grayscale information alone. Another unique element is that the residual fitting error is used to monitor tracking errors and outlier measurements, and can trigger the lip localization module for automatic re-initialization. We have implemented a real-time tracking system on a 195MHz SGI O2 workstation that runs at 30fps. Figure 1 shows a few tracking examples.

3. Hybrid Visual Features.

Hybrid features are comprised of both geometric- and pixel-based features. Using tracking results obtained from the algorithm described above, geometric-based features, including the width and height of the mouth area and their temporal derivatives, can be estimated automatically. Pixel-based features are derived from the vertical intensity profile calculated based on a subset of the pixels, delimited by the boundary of the upper and lower lips explicitly estimated by the tracking algorithm. The number of pixels that defines the profile varies over time as the lips open and close. By proper sub-sampling and linear interpolation, we map the vertical profile to a feature vector of constant length (e.g., 32 in our experiments). Therefore, information about the height of the mouth is largely decoupled from the pixel-based features. This is in contrast to cropping a rectangular region in the image that encompasses the lips in a sequence of image frames in an utterance, and subsequently taking the central vertical profile as the ROI. In practice, the ROI consists of a thin strip of pixels, where smoothing in the orthogonal direction is performed.

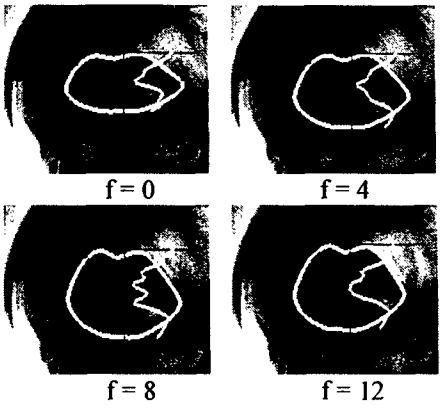

Figure 1: Snapshots of output from our lip tracking and visual feature extraction system in a few video frames. Geometric-based features were extracted from the tracking contour. Normalized pixel-based features were calculated based on the vertical intensity profile in the middle mouth region (plotted horizontally in light blue against a vertical axis).

Robustness of ROI estimation for pixel-based features and the accuracy of tracking are known to be important for improving accuracy of visual speech recognition [9][13]. The approach we proposed could also be applied to the whole ROI defined by the tracking contour as opposed to only to the vertical profile. Furthermore, transform-based features similar to that in [15] could also be derived and used as features instead. Comparison with these variants will be a subject of future study. In our experiments, the center profile contained much of the information about the appearance of the teeth and tongue, as well as their spatial relationship, and good recognition accuracy was achievable even in visual-only speech recognition.

Figure 1 illustrates the application of the tracking algorithm for the extraction of visual features (both geometric- and pixel-based).

4. HMM for Audio-Visual Speech.

Here we describe the basic elements of the HMMs in our approach.

An N-state HMM is characterized by a state transition matrix, $\{a_{ij}\}, 1 \le i, j \le N$, and a set continuous observation density functions, one for each state, which can be written as a Gaussian mixture

$$b_j(o_t) = \sum_{m=1}^{M} c_{jm} G(o_t; \mu_{jm}, V_{jm}), \quad 1 \le j \le N,$$

where o_t is the observation vector at time t, c_{jm} is the mixture coefficient, G is a multi-variate Gaussian distribution with mean μ_{jm} and covariance V_{jm} for mth mixture in the state j.

The acoustic and visual features were combined in two different ways in our HMM-based ASR experiments. In the first scheme, acoustic and visual feature vectors are concatenated to form individual feature vectors. In the second scheme, we model acoustic and visual features in separate feature streams. The mixture weights, mean vectors and covariance matrices in each observation density function are modeled separately in individual streams. The corresponding observation density is given by

$$b_i(o_t) = \left[\sum_{m=1}^{M_a} \alpha_{aim} G(o_{at}; \mu_{aim}, V_{aim})\right]^{\beta_a} \left[\sum_{m=1}^{M_v} \alpha_{vim} G(o_{vt}; \mu_{vim}, V_{vim})\right]^{\beta_v}$$

where subscripts a and v are used to denote the audio and visual channels, and the density of each channel is weighted by exponents β_a and β_v respectively, where $\beta_a + \beta_v = 1$. This is the multi-stream HMM formulation. The implicit assumption is that the audio and video observations are independent, which is really not exactly accurate. However, to be able to estimate reliably the parameters of b_i from limited amount of training data, it is customary to assume a diagonal covariance, and hence the assumption can be applied justifiably at least in the single Gaussian case with equal stream weights. Empirically, the stream weights can be used to give different emphasis to the observations, for example, based on the relative reliability of each channel.

5. Speech Recognition Experiments.

We performed a few evaluation experiments to compare various visual feature choices and investigate the relative merits of the various possible feature combinations. We focused on the connected digit recognition task. The eleven digits were 0-9 and 'oh.' The digit strings were taken from TIDIGITS, where utterances of up to seven digits were used. From a small database of 1518 audio-visual speech utterances, 759 were used for training and 759 for testing. Speech samples from one speaker were used to isolate the effects of speaker variability in this particular study. We used Hidden Markov Models to build word-model based recognizers. Gaussian mixtures were used to model the observation densities. The optimal number of mixtures (1-10) and number of hidden states (5-10) in the HMMs were determined empirically. A 3-state silence model was also used. The acoustic features were 12 Mel frequency cepstral coefficients (MFCC) plus the 0^{th} order cepstral coefficient, as well as their first and second temporal derivatives, resulting in an acoustic feature vector of size 39. They were computed every 10ms using a 25ms frame analysis window. Per-utterance cepstral mean normalization was also applied.

Table 1: Visual-only connected digit ASR's word error rate (WER %) for geometric (G), pixel-based (P), and hybrid (G+P) features described in this paper. The second and third rows are results with delta and delta-delta features. The size of the base feature vector is indicated in parentheses.

	G (2)	P(32)	G(2)+P(32)
Static	36.89	22.66	20.29
Static+Δ	26.88	11.59	9.88
Static+Δ+ΔΔ	27.80	9.49	**6.85**

The geometric features were derived from the width and height of the mouth normalized with respect to the corresponding dimensions when the speaker's mouth was closed. The pixel-based features were also normalized with respect to the mean value of the vertical profile when the speaker's mouth was closed. Interpolation of visual features was performed to generate samples at the audio feature frame rate of 100Hz.

In the audio-visual experiments, the audio features and visual features were concatenated to form a single feature vector for the single stream HMM case. The 2-stream HMM was also considered where the stream exponents were optimized using a linear step search. Alternatively, they could be discriminatively trained [17]. The Baum-Welch algorithm was used for EM-style embedded HMM training, and the Viterbi decoding algorithm for recognition. The HTK Toolkit [19] was used to design these experiments.

Table 1 shows first a summary of the recognition experiments employing visual features alone. One general trend we observed was that dynamic features (delta and delta-delta) in general carry additional information for

Table 2: Recognition WER (%) for the audio-only baseline (A), visual-only baseline (V), single stream audio-visual (AV1), 2-stream audio-visual (AV2) ASR at different SNR levels (dB). The reference visual feature used here was G+ΔG+P. β_a is the optimal stream weight on the audio channel for AV2. Note that AV1 was worse than the visual-only ASR at -3dB, whereas AV2 remained better.

	clean	20	15	10	5	3	0	-3
A	0.13	0.66	5.53	23.58	67.19	75.63	80.11	85.11
V	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26
AV1	0.13	0.53	1.32	2.50	7.38	10.14	15.55	22.79
AV2	0.13	0.26	0.53	2.50	6.59	9.75	12.12	14.49
β_a	0.95	0.85	0.8	0.65	0.5	0.45	0.35	0.35

recognition. Visual-only ASR word error rate as good as 6.85% was achieved, which was remarkable since no acoustic information was used and the pixel-based features were derived only from a small subset of pixels.

In the second experiment, we evaluated the effectiveness of the hybrid feature in the context of audio-visual speech recognition in the presence of noise. To be consistent with the visual features used in our previous work [4], the hybrid features employed were the combination of the base static pixel-based features, and the width and height of the mouth together with their first temporal derivatives (i.e., G+ΔG+P). We added F-16 cockpit noise (from the NoiseX database) to the audio channel systematically at various SNR levels (20dB to –3dB) only to the testing data. Table 2 summarizes the results. We observe that the bimodal recognizers consistently outperformed the audio-only counterpart at all SNR levels. Furthermore, the 2-stream HMM outperformed the single-stream HMM, and the performance difference increased as the SNR decreased. That was possible because the 2-stream HMM allowed stream weights to be applied selectively based on reliability of the acoustic features. In fact, the optimal stream weight on the audio channel decreased monotonically with the SNR level. We expect the overall performance will be higher if we use all delta and delta-delta visual features.

Figure 2 shows a screenshot of the tracking and audio-visual ASR system prototype that we have developed for experimentation.

6. Conclusion.

We overviewed a real-time visual lip tracking system that we used to define the ROI for visual feature calculation. We demonstrated the efficacy of our hybrid visual features in the context of connected digit recognition. Although single stream audio-visual HMM using concatenated features outperformed the acoustic-only counterpart, the 2-stream HMM gave the lowest WER at all SNR levels. The optimal stream weight for the audio channel decreased as the SNR level was lowered.

References

[1] C. Bregler and Y. Konig, "'Eigenlips' for robust speech recognition," in Proc. *International Conference on Acoustics Speech and Signal Processing*, pp. 669-672, 1994.

[2] M. T. Chan, Y. Zhang, and T. S. Huang, "Real-time lip tracking and bimodal continuous speech recognition," in Proc. *IEEE Signal Processing Society 1998 Workshop on Multimedia Signal Processing*, pp. 65-70, 1998.

[3] M. T. Chan, "Automatic lip model extraction for constrained contour-based tracking," in Proc. *IEEE International Conference on Image Processing*, Vol. 2, pp. 848-851, 1999.

[4] M. T. Chan: "HMM-based audio-visual speech recognition integrating geometric- and appearance-based visual features." In Proc. *IEEE Workshop on Multimedia Signal Processing*, pp. 9-14, Cannes, France, Oct 3-5, 2001.

[5] T. Chen, Rao, R. R., "Audio-visual integration in multimodal communication," in Proceedings of the *IEEE*, Vol. 86, pp. 837 –852, 1998.

[6] S. Chu, T. S. Huang. "Audio-visual speech modeling using coupled hidden Markov models," In Proc. *ICASSP*, 2002.

[7] S. Gurbuz, Z. Tufekci, E. Patterson, J. Gowdy, "Multi-stream product modal audio-visual integration strategy for robust adaptive speech recognition," In Proc. *ICASSP*, 2002.

[8] M. E. Hennecke, D. G. Stork, and K. V. Prasad, "Visionary speech: looking ahead to practical speechreading systems," in D.G. Stork and M.E. Hennecke (eds.), *Speechreading by Humans and Machines: Models Systems and Applications*, Springer, 1995.

[9] G. Iyengar, G. Potamianos, C. Neti, T. Faruquie, and A. Verma, "Robust detection of visual ROI for automatic speechreading," *Proc. IEEE Worshop on Multimedia Signal Processing*, Cannes, 2001.

[10] M. Kass, A. Witkin, and D. Terzopoulus, "Snakes: Active Contour Models," *International Journal of Computer Vision*, vol. 1, pp. 321-331, 1987.

[11] R. Kaucic and A. Blake, "Accurate, Real-Time, Unadorned Lip Tracking," in Proc. *6th International Conference on Computer Vision*, pp. 370-375, 1998.

[12] J. Luettin, N.A. Thacker, and S.W. Beet, "Visual speech recognition using active shape models and hidden Markov models," in Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, v 2, pp. 817-820, 1996.

[13] I. Matthews, G. Potamianos, C. Neti, J. Luettin, "A comparison of model and transform-based visual features for audio-visual LVCSR." In Proc. *International Conference on Multimedia Expo*, 2001.

[14] S. Nakamura, K. Kumatani, S. Tamura, "Robust bi-modal speech recognition based on state synchronous modeling and stream weight optimization," In Proc. *ICASSP, 2002.*

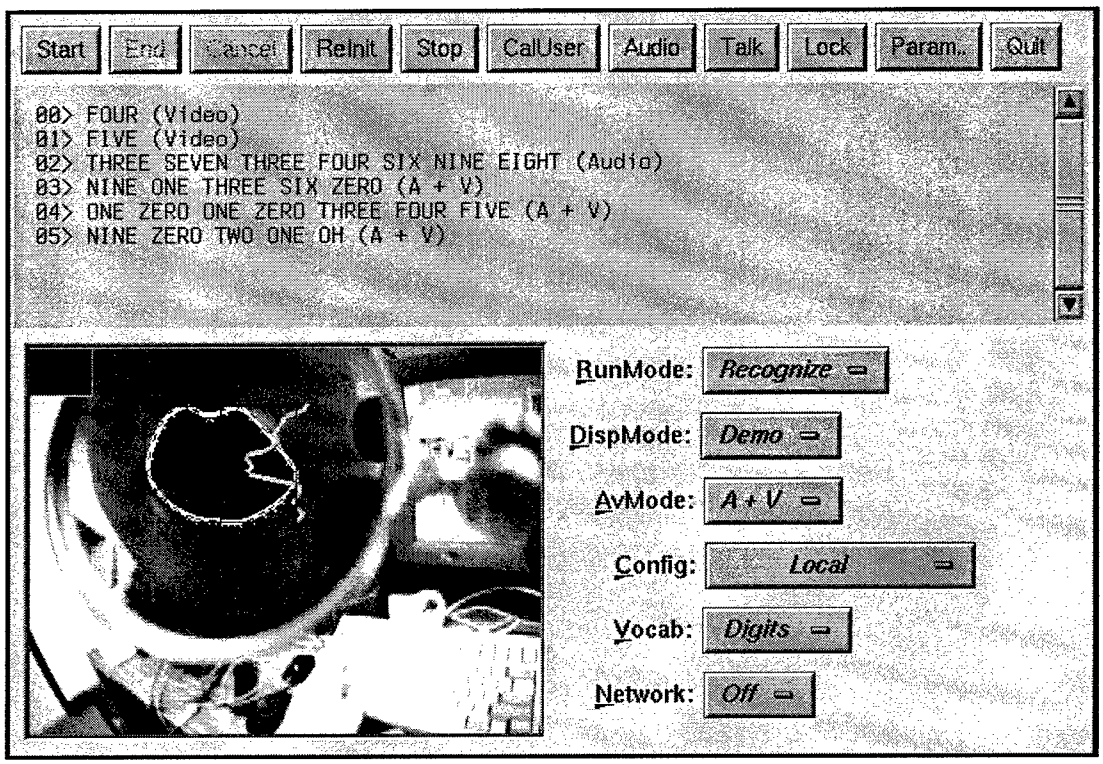

Figure 2: A screenshot of an experimental tracking and audio-visual ASR system at Rockwell Scientific. The system allows online switching among three recognition modes: audio-only, visual-only, or audio-visual. It can also be used to collect synchronized audio-visual sample data at 30fps directly to a disk array. A lightweight head-worn audio-visual capture apparatus can also be employed to allow users the freedom of head movement.

[15] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, "Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop," In Proc. *IEEE Workshop on Multimedia Signal Processing*, Cannes, 2001.

[16] E. D. Petajan, B. Bischoff, and D. Bodoff, "An improved automatic lipreading system to enhance speech recognition," in *ACM SIGCHI-88*, pp. 19-25, 1988.

[17] G. Potamianos, C. Neti. "Stream confidence estimation for audio-visual speech recognition," in Proc. *ICSLP*, vol III, pp. 746-749, 2000.

[18] R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition,* Prentice-Hall, New Jersey, 1993.

[19] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK-Hidden Markov Model Toolkit V2.1*, Entropic Research, Cambridge, 1997.

[20] A. L. Yuille, P. Hallinan, and D. S. Cohen, "Feature Extraction from Faces Using Deformable Templates," *International Journal of Computer Vision*, vol. 1, pp. 99-112, 1992.

[21] Y. Zhang, S. Levinson, and T. Huang, "Speaker independent audio-visual speech recognition," in Proc. *International Conference on Multimedia and Expo*, Vol 2, pp. 1073-6, 2000.